DOCUMENT RESUME

ED 144 950 TM 006 236

AUTHOR Hisama, Kay K.; And Others
TITLE Predictive Validity of Short Form Placement Tests under Two Scoring Systems.
PUB DATE [Apr 77]
NOTE 18p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, New York, April 5-7, 1977) ; Best copy available

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS College Placement; College Students; *Complexity Level; English (Second Language); Foreign Students; Higher Education; *Predictive Validity; Reading Tests; *Scoring Formulas; *Test Items; Test Results; *Test Validity
IDENTIFIERS Monte Carlo Methods; *Omitted Responses (Tests); *Test Length

ABSTRACT

The optimal test length, using predictive validity as a criterion, depends on two major conditions: the appropriate item-difficulty rather than the total number of items, and the method used in scoring the test. These conclusions were reached when responses to a 100-item multi-level test of reading comprehension from 136 non-native speakers of English were examined. Problems associated with scoring unanswered test items are discussed; the Monte Carlo method, in which responses are randomly assigned for unmarked items, was found to be more suitable than the raw scoring, or number correct method. (Author/MV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED144950

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PREDICTIVE VALIDITY OF SHORT FORM PLACEMENT TESTS UNDER TWO SCORING SYSTEMS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Kay K. Hisama

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Kay K. Hisama, Michael L. Bartlett, and Dennis W. Leitner

Southern Illinois University at Carbondale

Carbondale, Illinois

BEST COPY AVAILABLE

TM006 236

A paper presented at the annual meeting at the National Council on Measurement in Education
New York, 1977

PREDICTIVE VALIDITY OF SHORT FORM TESTS UNDER TWO SCORING SYSTEMS

Kay K. Hisama, Michael L. Bartlett, and Dennis W. Leitner

Southern Illinois University

This study examined the predictive validity of various short-form tests based on the responses to a 100 item multi-level test of reading comprehension, obtained from 136 non-native speakers of English. The results indicated that the optimal test length (predictive validity as a criterion) depends on two major conditions: (1) the appropriate item-difficulty rather than the total number of items, and (2) the Monte Carlo scroing of unmarked items rather than raw scoring (number right method).

INTRODUCTION

Test Length

"One of the best known properties of the tests commonly used in educational and psychological measurement (i.e., those composed of tasks which sample some particular domain of tasks) is that the longer they are (i.e., the larger the sample of tasks), the more reliable are the scores they yield (Ebel, 1972, p. 249)." This principle of educational measurement is reflected by the well known Spearman-Brown formula. The problem of this principle is that lay readers and even some professionals get the impression that the longer the test, the more reliable, therefore, more valid the test would be. Or conversely the shorter the test, the less reliable, and less valid. However, through their observation, practitioners of measurement are aware of the fact that lengthening a test is not a guaranteed way of increasing its reliability nor validity of a test. By the same token, shortening of a test may not necessarily reduce its reliability or validity.

The test length is an important factor in test construction because it affects the efficiency of a test, the important criterion of a good test. Nevertheless, studies of this sort are severly limited. Guilford et. al. (1942), for example, examined the problem only as a subsidiary question to their study of scoring systems: weighted and unweighted of a single multiple-choice test. In particular, they examined reliability and relative validity coefficients of three subtests, which were composed from the first 20 items, the first 50 items and 100 itmes respectively.

Reliability was determined by scoring each of the three tests of different length in both halves (odds & evens), and estimating the reliability coefficients by the Spearman-Brown formula in each case. The relative validity was determined by correlating each

of three subtests and the total test scores (the sum of 201 multiple-choice and 107 true-false items). While the coefficients of reliability of scores increased substantially under the two scoring systems, the validity coefficients increased only slightly as the number of items increased. The major criticism of Guilford's study is that it does not have an outside criterion, thus the correlation is spurious. In addition, the lack of outside criterion makes it difficult to interpret the results.

After Guilford, a number of authors discussed the problem of test length in terms of test reliability or its variation (Lord, 1957, 1959; Ebel, 1972, Cureton et.al., 1973), but little attention was given to the validity per se. One of the primary purposes of this investigation was to determine the effects of various test lengths on some meaningful external criteria. The other important problem of this study is the problem of omitted items under two scoring systems.

The Problem of Omitted Items

A number of authors have noted the prevalence and the resulting problem of omitted items. Although a significant number of omissions were observed even with mature students, (Lord, 1964) the problem seems to be profound among unsophisticated subjects. In a study involving 1,354 ninth grade students, Raffeld (1975) noted "...so many omissions were present in the Oregon sample (p. 180)." In another study, Sabers and Feldt (1968) felt that their results indicated the regular instructions "were not forceful enough to encourage all students to attempt every item (p. 256)." In still another study, Kansup and Hakstian (1975) had to eliminate 10% of their 1029 grade nine students who served as subjects, because they left one or more itmes unmarked even though they were instructed to answer every item.

In spite of its potential impact on various aspects of testing, the problem of omitted items has never been investigated seriously. It should be noted that most researchers treated the problem as a part of the problem of guessing even though the two are different. Kansup and Hakstian's study is a rare example in which the problem of omitted items was isolated from that of guessing. Focusing on the guessing problem, they simply dropped those subjects who left one or more items unmarked.

Moreover, the researchers who pursued the problem of guessing focused on test directions as the single most important factor although there are a number of other factors which would influence the number of omissions, i.e., the level of sophistication, achievement motivation, and cultural background of the subjects. Lord defined two commonly used test directions: number-right directions and formula-scoring directions. Number right directions will inform the examinee convincingly that answering every item is his best strategy. If there is a scoring penalty for wrong answers, on the other hand, the examinee should be informed as to how he can maximize his expected score--formula scoring directions (Lord, 1975). However, these test directions do not state what to do if the time runs out.

Davis maintains that an item may not be marked for two quite different reasons: (1) after reading the item, the examinee realizes he does not know the answer and does not guess the right answer, and (2) the examinee does not have time to consider the items (Davis, 1951). Yet no empirical studies discussed these differences explicitly. This study investigated the scoring of unmarked items using a power test which consists of well graded items, thus focusing on the first of the two possibilities indicated by Davis.

Test Instrument

The test investigated was the Reading for Understanding Placement Test (RFU), developed by Thelma G. Thurston (1959) as a part of the Reading for Understanding program. The program purports to improve the reading comprehension of students from fifth grade through high school and in a college reading clinic. The test contains 100 four-choice items in ascending difficulty, spaced at equal intervals of scale value. The test instrument is one of the placement test used to place the subjects of this study into one of six levels.[1]

Method

Subjects

Subjects were 136 foreign students at the Center for English as a Second Language (CESL) at Southern Illinois University at Carbondale. There are eight sessions in a year and a total of 50 to 100 students take the placement battery at the beginning of each session. Out of eight sessions in a year, the following sessions were sampled: October, November, and January. Two criteria for the subject were that he or she had to be a new entering student to the center, and that he or she had not been in the United States longer than one month at the time of testing.

Test Administration

Subjects took the test as a part of a placement battery upon entrance to CESL. The test was the last one of the three given consecutively. The general directions for the first two

---

[1]The six course levels vary from Course I -- Knows a little English; can use only the basic structure --, Course V-- can converse fluently on most subjects to Course VI, a course specifically designed for graduate students.

tests[2] were as follows.

> Mark only one answer for each problem. If you change your mind about an answer after you have marked it on the answer sheet, completely erase your first answer and then mark your new answer.
>
> Try to answer every problem. If you are not sure of the correct answer, make the best guess that you can. Your score on the test will be based on the number of correct answers that you give.
>
> (Harris and Palmer, 1970)

Upon the administration of the RFU, students were told to do just as they did in taking the two previous ones. Therefore, the test directions for the RFU are the same as "number-right scoring directions" in Lord's definition (Lord, 1975). The RFU is a power test and Thurston suggests a practical time limit of 50 minutes. However, CESL students were given 75 minutes which was found appropriate because beyond this time limit the unattempted problems become too difficult to the group. All tests were given by the first author under identical test directions. The responses to the test were obtained using one of the standard IBM answer sheets.

Scoring

Two scoring systems were applied: (a) the number-right scoring and (b) Monte Carlo scoring of the unmarked items. In the first method, the scores were obtained by counting the number of right answers. The second method requires a further explanation. The basic idea of Monte Carlo scoring applied in this study was to assign random responses by computer for all items omitted consecutively to the last item. The beginning of the omitted items was defined as two or more blanks which were being followed by no consecutively marked items. Then, using the appropriate probability distribution,[3] the number of items for which examinees would have been given credit if they were forced to mark all the items, were determined.

Rationale for this scoring is as follows. Upon the examination of each answer sheet, it was found that examinees, with a few exceptions, left blank consecutive items starting somewhere in the test to the last item; a pattern which seems to indicate they did not have time to consider those items. However, the following observations seemed to justify that the distribution of these omitted responses would, if they were forced to mark all

---

[2]The two tests were: A Comprehensive English Language Test for Speakers of English as a Second Language (CELT) for Structure--Form A and the same test for Listening--Form A.

[3]See appendix B of this paper.

the items, tend to be more like a chance distribution than any other: (1) the examinees were given 75 minutes, 50% more time than was recommended in the test manual, (2) items have been carefully calibrated and arranged in order of difficulty, and (3) a large number of students indicated that the remaining items were too difficult.

Scores for different test lengths were defined by scoring various lengths of short-form tests based on the responses to a total of 100 items. In particular, various short-form tests were defined in the following fashion. (1) Divide 100 items into ten items. In particular, various short-form tests were defined in the following fashion. (1) Divide 100 items into ten blocks; B_1, B_2, ...B_{10}, where:

B_1 contains items #1 through #10
B_2 contains items #11 through #20
.
.
.
B_{10} contains items #91 through #100.

(2) Add the scores of the desired number of consecutive blocks of items in order to find out scores on various short-form tests.

In defining short-form tests, somewhat different approaches were taken for the two scoring systems. In the case of raw scoring, the scores on the short-form were determined by adding three or more consecutive blocks, but always starting from the first block. This seemed to be an appropriate procedure because there is a strong possibility that the student would have marked additional items if some items in the preceding blocks were dropped. For Monte Carlo scoring, three of more of consecutive blocks of any range were added.

Method of Analyses

The analysis consists of a systematic comparison of the coefficients of predictive validity. They were obtained by correlating the criterion and predictor variables defined as follows. Criterion variables: All CESL teachers who have one or more new students in their classes were asked to evaluate them at the end of one CESL session in two respects,[4] (1) the recommended level of reading class for the next session, and (2) the relative class standing of each student (High-Middle-Low).

---

[4]Teacher evaluation was made on the basis of classroom examinations compiled by CESL teachers. Examination questions, which are of multiple choice form, covered the reading materials used in the classroom, and eas uniform throughout different classes at the same level.

On the basis of these information two criterion variables were defined.

LEVEL - the appropriate level of reading class at CESL in which the student will be placed in the following session. There were seven levels for this purpose because the abilities of some students were above the highest of six classified levels at CESL.

RANK -- a modified version of LEVEL by incorporating the student's relative class standing. There were thirteen ranks.

Predictor variables: Predictor variables are test scores on eight different lengths of the RFU under the two scoring systems. Short-form tests were determined by truncating one or both ends from a total of 100 test items. Two types of scores were as follows.

RAW SCORES -- scores on the RFU based on number-right scoring

ADJUSTED SCORES -- scores on the RFU which are the sums of the raw scores and the credit obtained for Unmarked items using Monte Carlo scoring.

RESULTS

Table 1 shows validity coeffieients for full and seven short-form tests under the number-right scoring method. The test was shortened by truncating only the upper end of the full length test.

TABLE 1

VALIDITY COEFFICIENTS FOR FULL AND SHORT-FORM TESTS BASED ON RAW SCORES

	Number of Items*							
Criterion	30	40	50	60	70	80	90	100
LEVEL	.712	.765	.780	.802	.802	.813	.813	.816
RANK	.731	.785	.803	.824	.825	.833	.831	.834

*The full-length test contains 100 items in ascending difficulty. The various test lengths were defined by cutting the upper end of test itmes from the full length test.

These coefficients demonstrate that the increase in validity coefficients is very small beyond 60 itmes. Criterion RANK

show higher validity than those of LEVEL indicating that there are considerable differences in their proficiency in reading ability among those who are placed in the same level.

Table 2 shows validity coefficients under Monte Carlo scoring of unmarked items. This table parallels Table 1.

TABLE 2

VALIDITY COEFFICIENTS FOR FULL AND SHORT-FORM TESTS BASED ON RAW SCORES

	Number of Items*							
Criterion	30	40	50	60	70	80	90	100
LEVEL	.729	.786	.800	.817	.821	.830	.825	.821
RANK	.759	.808	.825	.842	.846	.852	.846	.844

*Same as Table 1

An examination of Table 2 indicates that there is a close agreement in the general trend in the two tables: (1) criterion RANK shows higher validity coefficients than those of LEVEL, and (2) the increase in validity coefficients levels of beyond 60 items. Table 3 is somewhat different from the above two tables in that ranges are defined by truncating one or both ends from the full length test.

TABLE 3

VALIDITY COEFFICIENTS FOR VARIOUS TEST LENGTH BASED ON ADJUSTED SCORES N=136

3 - A

Criterion	01-40	Range of test items K=40 11-50	21-60	31-70
LEVEL	.786	.810	.823	.779
RANK	.808	.834	.846	.805

3 - B

Criterion	01-50	Range of test items K=50 11-60	21-70	31-80
LEVEL	.800	.821	.820	.779
RANK	.825	.844	.845	.798

3 - C

Criterion	01-60	Range of test items K=60 11-70	21-80	31-90
LEVEL	.817	.823	.825	.766
RANK	.842	.848	.845	.785

3 - D

Criterion	01-70	Range of test items K-70 11-80	21-90	31-100
LEVEL	.821	.831	.816	.761
RANK	.846	.851	.836	.871

The test contains 100 items in ascending difficulty. The ranges of K items were defined by truncating one or both ends from the full length test.

From Table 3, it appears that increasing the number of items does not necessarily increase the test validity, but choosing items of optimal difficulty does. For example, when LEVEL is used as a criterion, the highest validity coefficient of 40 items test (r_{val} = .823) is higher than three out of four validity coefficients of 70 items tests (r_{val}'s are .821, .831, .816, and .761) It follows, then, that it may not be necessary to administer all test items of this multi-level test for the purpose of placing new CESL students, thus producing a considerable saving in testing time. As in Tables 1 and 2, higher validity coefficients are obrained for the criterion RANK than for LEVEL.

Turning to the scoring method, the data indicates that counting the number of right answers is not the best approach because many examinees left a large number of items unmarked. Therefore, an adjusted score was derived for each examinee. The validity coefficients of the adjusted score are consistently higher than those of raw scores regardless the differences in test length or criterion variable. The difference in the respective validity coefficients under the two scoring was about .02. The amount of increase is impressive in reference to a point made by Lord. Lord (1975) has stated that if a new scoring method increases test reliability from .90 to .91 without extra cost, it is equivalent to an increase of about one-eighth in test length. In other word, if we do not need the increased reliability, we can use the new scoring method and shorten the test by one-ninth. In the case of validity, the effect is more dramatic. If the old scoring method gave a validity of .6, the new method will (by standard fromulas) give a validity of .6033.If we do not need the increased validity, we can use the new scoring method and shorten the test by one-ninth without loss of validity. If one compares that the increase of .02 obtained for this study with .0033, the value used by Lord, superiority of Monte Carlo scoring over the number-right scoring is evident.

It should be noted that this study had an large number of unsophisticated subjects who left many items unmarked, thus were evaluated unfairly. In other research, the subject who leaves a large number of items may be a minority member. Then it will be difficult to demonstrate the difference which may exist among different scoring systems.

DISCUSSION

This study investigated two important aspects of educational measurement, test length and the scoring of omitted items, for which little empirical studies have been made. Concerning the test length, the data indicated the importance of item difficulty, rather than the total number of items, for the group for which the test was used. The predictive validity of only 40 items was as high as those of 70, 80, 90, or 100 items. This indicates that the testing time could be shortened by as much as one-half. The reduction in the number of test

items would most benefit test consumers: the shorter time for taking a test at a lower cost.

As to the scoring of omitted items, consistently higher validity coefficients obtained under the Monte Carlo scoring of unmarked itmes than the number-right scoring seems to warrant that it is important to study this problem. The following factors seem to have contributed to demonstrate such results: (1) there were a large number of unsophisticated subjects who were likely to leave test items unmarked, and (2) the test consists of a large number of well graded items including those which were too difficult for the group. In the case of minority students, who also are likely to leave test items unmarked, it will be difficult to demonstrate the effect of omitted items in terms of group statistics such as validity coefficient because of the small number of such students in the group. Since most multiple-choice tests are scored by the number-right method, subjects who omit the items are likely to be under-scored.

We all know only a few points on a test often have significant bearing in determining a student's success or failure. At CESL, for example, missing a few points often results in being placed in one level lower than where he could have been with an additional few points. This means the student had to delay at least six week in entering a college at a minimum cost of $500. The length of delay might be as long as one academic term. This is a serious problem and requires a further discussion.

Various suggestions have been made in order to reduce the impact of omitted items. Lord, a long time advocate of "formula scoring", the procedure generally known as "the correction for guessing" with a direction which discourages random guessing, recently restated his position (Lord, 1975). A major problem of formula scoring is that even though one agrees with Lord's basic assumptions of formula scoring, there would be much disagreements as to how one should go correcting the score. Correction methods which count the numbers of right and wrong answers would be too cumbersome to be used by classroom teachers.

Lord is right in stating "if there are any omitted items at all under formula scoring, the number-right score $\underline{x}$ is an 'inadmissible' estimator,..." However, he seems to have placed too much emphasis on test directions. For many examinees, leaving items unmarked seems to be a matter of acquired habit in taking multiple choice tests, rather than the consequence of a sound judgement made from time to time on the basis of specific test directions the assumption which underlies Lord's formula scoring. The results of this study indicated that some kind of adjustment of scores is necessary regardless of different test directions as long as there are any omitted items.

Mollenkopf, (1960) has suggested somewhat different approach to scoring omitted items. His preference is for instructing the candidate that he will automatically receive u/k points of credit for items he has not reached where $\underline{U}$ represents the number of items not attempted and $\underline{k}$ is the number of alternatives per

item. Monte Carlo scoring of unmarked items adopted in this study, is somewhat different from Mollenkopf's approach, but the two are motivated by the same principles: (1) to bring the chance element in test scores under some control, and (2) allow the timid to profit equally with the bold. The results of this study have demonstrated that this line of approach in correcting for guessing does indeed increase the test validity. Advantages of Monte Carlo scoring over Mollenkopf's are that there will be no cumbersome computations nor decimal numbers, which might introduce error if the procedure is applied routinely by classroom teachers. A simple table is to be developed which can be used by classroom teachers in order to simplify Monte Carlo scoring.

CONCLUSIONS

This study focused on two important, but neglected aspects of educational measurement. Quite dramatic effects of two variables, test length and scoring of omitted items, on test validity demonstrated in this study seem to warrant the needs for further investigation.

Measurement specialists have paid too much attention to the aspect of reliability, especially its mathematical and statistical aspects based on the true score model, in spite of much criticism such as a "grinding of mathematical symbols to produce nothing of any worth (Lumsden, 1975, p.257)."

It is hoped that the findings of this study would generate much enthusiasm in reasearch motivated to improve the validity and efficiency of tests, administered at institutional levels. Such efforts on the part of measurement specialists would definetely benefit test consumers.

REFERENCES

Cureton, E.E. Length of test and standard error of measurement. Educational and Psychological Measurement, 1973, 33, 63-68.

Davis, F.B. Item selection techniques, In E.F. Lindquist, (Ed.) Educational Measurement, Washington, D.C." American Council on Education, 1951.

Ebel, R.L. Why is a longer test usually a more reliable test? Educational and Psychological Measurement, 1972,32, 249-253.

Guilford, J.P., Lovell, C., Williams, R.M. Completely weighted versus unweighted scoring in an achievement examination. Educational and Psychological Measurement, 1942, 2, 15-21.

Harris, D.P. & Palmer, L.A. A comprehensive English Language Test for Speakers of English as a Second Language: Test Booklet. New York: McGraw-Hill, 1970.

Kansup, W. & Hakstian, A.R. A comparison of several methods of assessing partial knowledge in multiple choice tests: I. Scoring procedures, Journal of Educational Measurement, 1975, 12, 219-230.

Lord, F.M. Do tests of the same length have the same standard errors of measurement? Educational and Psychological Measurement, 1957, 17, 510-521.

Lord, F.M. Tests of the same length do have the same standard error of measurement. Educational & Psychological Measurement, 1959, 19, 233-239.

Lord, F.M. The effect of random guessing on test validity. Educational and Psychological Measurement, 1964, 24, 745-747.

Lord, R.M. Formula scoring and number-right scoring. Journal of Educational Measurement, 1975, 1, 7-11.

Lumsden, J. Test theory, In M.R. Rosenzweig and W. Porter (ed.s) Annual Review of Psychology, 1976, 251-280.

Mollenkopf, W.G. Time Limits and the Behavior of test takers. Educational and Psychological Measurement, 1960, 20, 223-30.

Raffeld, P. The effects of Guttman weights on the reliability and predictive validity of objective tests when omissions are not differentially weighted. Journal of Educational Measurement, 1975, 12, 179-186.

Sabers, D.E & Feldt, L. an empirical study of the effect of the correction for chance success on the reliability and validity of an aptitude test. Journal of Educational Measurement, 1968, 5, 251-258.

Thurston, G.T. General Reading for Understanding, North Carolina: Science Research Associates, Inc., 1959.

APPENDIX A

MEANS AND STANDARD DEVIATIONS (SDs) OF TEST SCORES BASED ON 136 CESL STUDENTS

Ranges	N of items	Means	SD
01-30 *	30	22.93	6.54
01-40 *	40	28.68	8.67
01-50 *	50	33.88	10.85
01-60 *	60	37.42	12.66
01-70 *	70	40.12	13.77
01-80 *	80	41.90	14.60
01-90 *	90	43.28	15.20
01-100 *	100	44.37	15.78
01-30	30	23.09	6.10
01-40	40	28.93	8.05
01-50	50	34.21	10.07
01-60	60	37.97	11.63
01-70	70	41.09	12.28
01-80	80	43.43	12.71
01-90	90	45.73	12.81
01-100	100	48.08	12.86
01-40	40	28.93	8.04
11-50	40	25.62	8.44
21-60	40	21.15	8.65
31-70	40	18.00	7.23
01-50	50	34.21	10.07
11-60	50	29.38	10.09
21-70	50	24.26	9.37
31-80	50	20.35	7.82
01-60	60	37.97	11.63
11-70	60	32.50	10.76
21-80	60	26.61	9.87
31-90	60	22.64	7.99
01-70	70	41.09	12.28
11-80	70	34.85	11.22
21-90	70	28.90	10.01
31-100	70	24.99	8.03

Asterisk (*) indicates a raw score. Others are adjusted scores in which responses to unmarked items were simulated on the digital computer.

APPENDIX B

COMPUTER PROGRAM: SIMULATED RESPONSES

```
C  THIS PROGRAM GENERATES 100 RANDOM NUMBERS FOR EACH OF J SUBJECTS AND
C  EVALUATE THESE NUMBERS IN TERMS OF CREDIT (1), NO CREDIT (0) BASED ON
C  P, THE PROBABILITY OF GETTING RIGHT ANSWERS BY RANDOM GUESSING.

      DIMENSION X(100),N(100)
      DATA P/0.25/
      DATA JW/0/
      DATA JX/91767/
      DATA JZ/0/
      DO 4 J=1,136
      JV=0
      DO 3 K=1,100
    1 CALL RANDO(JX,JY,Y)
      JX=JY
      IF(Y.LE.1.0.AND.Y.GT.0.0)GOTO 2
      JZ=JZ+1
      WRITE(6,9)J,K,JZ,Y
      IF(JZ.LT.100)GOTO 1
      WRITE(6,8)
      CALL EXIT
    2 X(K)=Y
      N(K)=0
      IF(Y.GT.P)GOTO 3
      JV=JV+1
      N(K)=1
    3 CONTINUE
      WRITE(6,7)JV
      JW=JW+JV
      DO 4 K=1,4
      LA=K*25-24
      LZ=K*25
      WRITE(6,6)J,K,(S(L),L LA,LZ,(N(L),L LA,LZ)
    4 WRITE(7,5)J,K,(N(L),L LA,LZ)
      WRITE(6,44)JZ,JW
      CALL EXIT
   44 FORMAT(////I9,'ERRORS ENCOUNTERED'/I9,'ONE VALUES EMITTED')
    5 FORMAT(I4,I2,4X,25I1)
    6 FORMAT(I4,I2,4X,25(1X,F3.2)/9X,25I4)
    7 FORMAT('0',I9,"ONES EMITTED"/)
    8 FORMAT('****LIMIT OF INVALID RESULTS SURPASSED")
    9 FORMAT(' ',20('*'),'INVALID RESULT FOR SUBJECT',I4,'ITEM',I4,
     1' VALUE',E16.8)
      END
```

Requests for the entire program used for this study should be sent to Michael bartlett, Dept. of Guid. & Ed. Psych., Southern Illinois University, Carbondlae, Illinois 62901.

APPENDIX C

ITEM MEANS AND NUMBER OF UNMARKED ITEMS FOR RFU (N=136)

ITM	MEAN	UNMARKED	ITM	MEAN	UNMARKED	ITM	MEAN	UNMARKED
1	.913	4	41	.601	7	81	.203	47
2	.891	3	42	.406	14	82	.130	48
3	.870	5	43	.580	7	83	.101	49
4	.775	5	44	.775	7	84	.210	52
5	.928	4	45	.442	12	85	.210	55
6	.826	9	46	.471	11	86	.109	53
7	.906	6	47	.587	14	87	.196	58
8	.790	5	48	.261	20	88	.058	60
9	.783	5	49	.558	9	89	.116	60
10	.884	4	50	.565	9	90	.087	64
11	.717	5	51	.420	20	91	.080	63
12	.761	6	52	.312	24	92	.109	66
13	.763	6	53	.283	18	93	.159	66
14	.913	6	54	.333	23	94	.174	65
15	.920	4	55	.406	19	95	.138	66
16	.884	3	56	.370	21	96	.138	66
17	.928	5	57	.377	22	97	.109	70
18	.732	12	58	.261	18	98	.101	73
19	.732	5	59	.435	23	99	.072	69
20	.855	6	60	.413	22	100	.094	75
21	.355	14	61	.377	24			
22	.732	6	62	.290	28			
23	.652	8	63	.254	26			
24	.558	12	64	.304	25			
25	.580	6	65	.290	26			
26	.594	9	66	.312	30			
27	.543	12	67	.326	26			
28	.609	11	68	.203	29			
29	.855	7	69	.188	30			
30	.732	5	70	.181	33			
31	.486	16	71	.174	33			
32	.725	8	72	.181	39			
33	.529	10	73	.384	34			
34	.667	6	74	.181	33			
35	.457	8	75	.188	37			
36	.384	20	76	.188	42			
37	.667	9	77	.174	39			
38	.507	8	78	.087	37			
39	.768	6	79	.080	41			
40	.623	8	80		45			